PLOS | ONE

RESEARCH ARTICLE

CardioGenBase: A Literature Based Multi-Omics Database for Major Cardiovascular Diseases

Alexandar V[1], Pradeep G. Nayar[2], R. Murugesan[1], Beaulah Mary[1], Darshana P[1], Shiek S. S. J. Ahmed[3]*

1 Faculty of Allied Health Sciences, Chettinad Academy of Research and Education, Kelambakkam 603 103, Tamil Nadu, India, **2** Department of Cardiology, Chettinad Super Specialty Hospital, Chettinad Academy of Research and Education, Kelambakkam 603 103, Tamil Nadu, India, **3** Department of Computational Biology, Drug discovery Lab, Faculty of Allied Health Sciences, Chettinad Academy of Research and Education, Kelambakkam, 603 103, Tamil Nadu, India

* shiekssjahmed@gmail.com

OPEN ACCESS

Citation: V A, Nayar PG, Murugesan R, Mary B, P D, Ahmed SSSJ (2015) CardioGenBase: A Literature Based Multi-Omics Database for Major Cardiovascular Diseases. PLoS ONE 10(12): e0143188. doi:10.1371/journal.pone.0143188

Editor: Xia Li, Harbin Medical University, CHINA

Received: August 18, 2015

Accepted: November 1, 2015

Published: December 1, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information file.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Cardiovascular diseases (CVDs) account for high morbidity and mortality worldwide. Both, genetic and epigenetic factors are involved in the enumeration of various cardiovascular diseases. In recent years, a vast amount of multi-omics data are accumulated in the field of cardiovascular research, yet the understanding of key mechanistic aspects of CVDs remain uncovered. Hence, a comprehensive online resource tool is required to comprehend previous research findings and to draw novel methodology for understanding disease pathophysiology. Here, we have developed a literature-based database, CardioGenBase, collecting gene-disease association from Pubmed and MEDLINE. The database covers major cardiovascular diseases such as cerebrovascular disease, coronary artery disease (CAD), hypertensive heart disease, inflammatory heart disease, ischemic heart disease and rheumatic heart disease. It contains ~1,500 cardiovascular disease genes from ~2,4000 research articles. For each gene, literature evidence, ontology, pathways, single nucleotide polymorphism, protein-protein interaction network, normal gene expression, protein expressions in various body fluids and tissues are provided. In addition, tools like gene-disease association finder and gene expression finder are made available for the users with figures, tables, maps and venn diagram to fit their needs. To our knowledge, CardioGenBase is the only database to provide gene-disease association for above mentioned major cardiovascular diseases in a single portal. CardioGenBase is a vital online resource to support genome-wide analysis, genetic, epigenetic and pharmacological studies.

Introduction

Cardiovascular diseases are the leading cause of morbidity and mortality worldwide[1]. Among the cardiovascular conditions, cerebrovascular disease, coronary artery disease (CAD), hypertensive heart disease, inflammatory heart disease, ischemic heart disease and rheumatic

heart disease are considered as major cardiovascular diseases (MCVDs) that are caused by both genetic and epigenetic factors resulting in heart failure. The pathophysiology of MCVDs are not merely the result of single gene defect or its product alone. It is an outcome of several molecules, which function collaboratively to initiate oxidative stress, inflammation, endothelial dysfunction and thrombosis. To date, the polygenic nature of MCVDs is highly accepted[2,3]. Several studies have been conducted on MCVDs which includes association studies, linkage studies and meta-analyses that identified various diseases-associated genes[4–9]. These findings generated an unprecedented amount of biological data that provide an opportunity to construct a useful gene resource for MCVDs.

A broad knowledge of genes and proteins involved in cardiovascular conditions is crucial for understanding of molecular mechanism in disease pathology. Here, we present a comprehensive gene database (CardioGenBase) for the major cardiovascular diseases. The CardioGenBase (http://www.CardioGenBase.com/) is a knowledge base which effectively integrates, analyzes and visualizes major cardiovascular disease associated research articles. It was constructed by collecting gene/protein information across MCVDs related published literatures. The identified entities were enriched with chromosomal location, gene ontology, gene expression, protein expression, bioavailability, pathways, SNPs, protein interaction network and drugs. In addition, it enables users to search and browse various data categories and data connections. CardioGenBase is a unique genetic resource that would help cardiovascular research community to design new experiments and to unveil novel disease mechanisms.

Results and Discussion

CardioGenbase was created as literature evidence based database to provide useful molecular information on major cardiovascular diseases (Fig 1). The scientific literature was manually collected, filtered and a computer program (Lucene)was used to identify gene/protein names from the collected articles. Lucene is an open source and a java based computer program. It is effective

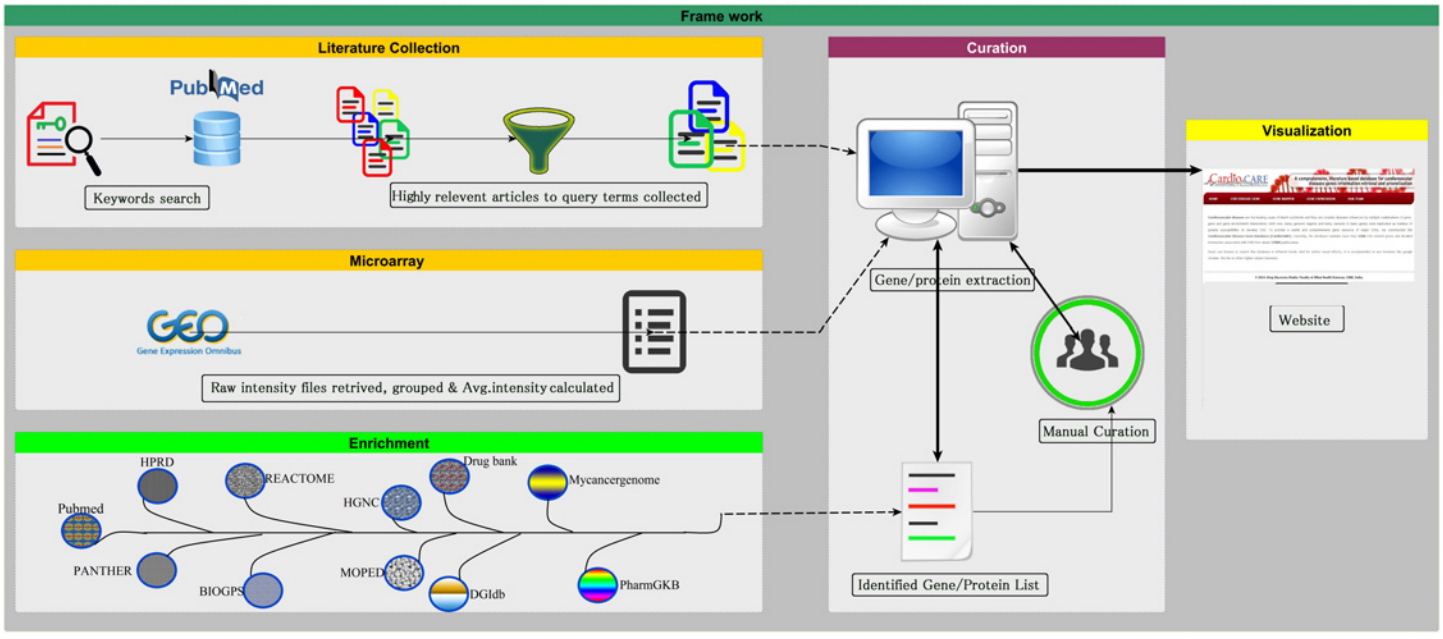

Fig 1. CardioGenBase Construction. The framework describes the construction of CardioGenBase. It includes data mining of biomolecules, filtration, curation, enrichment, system interface and visualization.

doi:10.1371/journal.pone.0143188.g001

for full-featured text mining. Using this program, we identified 1365 genes for CAD, 240, 75, 28, 428 and 139 for cerebrovascular disease, hypertensive heart disease, inflammatory heart disease, ischemic heart disease and rheumatic heart disease, respectively (Table 1). The data obtained are categorized, stored and managed as tables using MySQL to create CardioGenbase.

The genes in the database were enriched with gene expression, protein expressions, ontology, SNP, PPI network, drugs and pathways. These molecular information is a prerequisite to design and conduct basic research to understand disease pathophysiology and to discover biomarker(s). Therefore, CardioGenBase contains both gene and protein expression profiles of more than 30 and 10 tissues, respectively. In addition, protein-protein interaction (PPI) networks and pathways are provided to understand disease molecular mechanism. Here, the PPI network shows the interaction of disease gene with other key molecules to execute a molecular function(s) through single/multiple pathways[10]. Further, all the associated pathways were given to show the involvement of the query gene in various molecular processes. Furthermore, user can magnify these pathway images in a new window for better perceptive, and those images can be downloaded. Also, the database consists of gene-drug information such as inhibitor, stimulator and suppressor which are helpful in pharmacological studies. All these data are organized into four different tools in the web interface.

Tool 1: Disease Finder

The *disease finder* provides genes that are associated to a major cardiovascular disease (Figure A in S1 File). User can select any cardiovascular disease of their interest from the list to retrieve complete genes of the selected cardiovascular disease. This tool enables the user to identify the reported genes for the given disease condition (Fig 2).

Tool 2: CVD Gene Finder

CVD gene finder allows the user to search for a gene to any major cardiovascular disease covered in the database (Figure B in S1 File). This tool aids the user to search earlier scientific reports on the query gene for the disease of interest (Fig 3). User needs to select an MCVD and the query gene (HGNC ID or official Gene Symbol). The results for the queried gene consists of literature evidences including abstract, Pubmed IDs and journal citation along with the detailed molecular information about the gene such as ontology, SNP, PPI network, pathways, drugs along with the literature evidences.

Tool 3: Gene Mapper

Gene Mapper helps the user to search multiple genes at once to identify its cardiovascular disease associated (Figure C in S1 File). The gene Mapper generates a Venn diagram that displays

Table 1. Text mining results. The number of literature collected for each cardiovascular disease. These literature was filtered based on title/abstract, relevance to the search terms to extract genes/proteins using a semi-automated method.

Disease	No. Literature	Genes Extracted
Cerebrovascular disease	1966	240
Coronary Heart Disease	17471	1365
Hypertensive heart disease	260	75
Inflammatory heart disease	23	28
Ischemic heart disease	5624	428
Rheumatic heart disease	644	139

doi:10.1371/journal.pone.0143188.t001

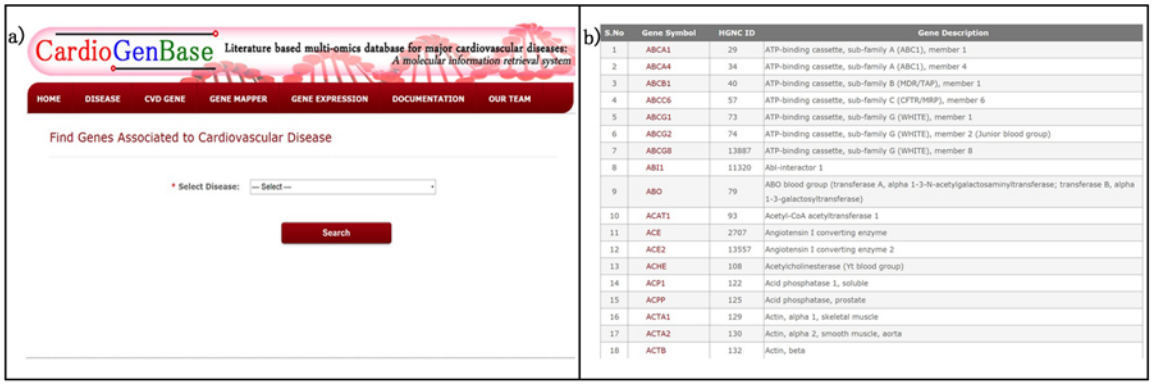

Fig 2. Disease Finder. a) All the reported genes associated a major cardiovascular disease could be retrieved using this query page. b) The result page showing all the genes associated with a disease of interest.

doi:10.1371/journal.pone.0143188.g002

user input gene list and number of cardiovascular associated genes from the input list (Fig 4). For each cardiovascular associated gene, the literature evidence was provided that enable the user to rank or prioritize the query genes based the given literature evidence.

Tool 4: Gene Expression Finder

Gene expression finder enables users to identify the expression of a gene under various cardiovascular disease conditions (Figure D in S1 File) The microarray gene expression data for cardiovascular disease were used retrieved from NCBI GEOSET. Here, the raw intensity of the samples are collected, grouped and the average intensities is displayed (Fig 5). This feature is similar to the NCBI GEO profile viewer[11], but specific to cardiovascular disease conditions. This tool enables the user to identify the differentially expressed genes in the selected experimental condition.

Comparison and Validation

To our knowledge, CardioGenBase is the only database that integrates six major cardiovascular conditions with gene to publication associations from ~24000 research articles. In order to evaluate the accuracy and credibility of CardioGenBase, the manually curated CADgene database[12] was used as a "gold standard" which was updated in the year 2013. For the fair comparison, the articles published between the years 1988 to2013 was used for the validation

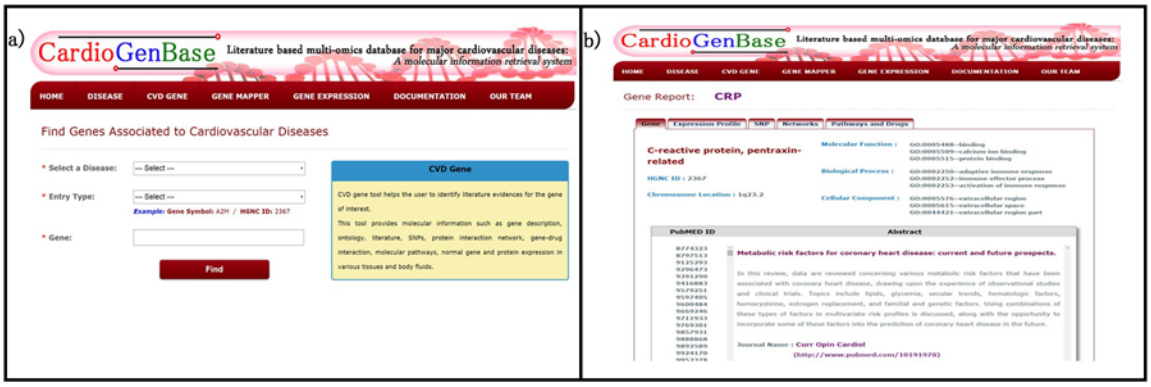

Fig 3. CVD Gene Finder. a) The literature evidence and molecular information could be obtained for a gene of interest. User can search the gene by HGNC ID or gene symbol. b) The output shows the molecular information on the query gene.

doi:10.1371/journal.pone.0143188.g003

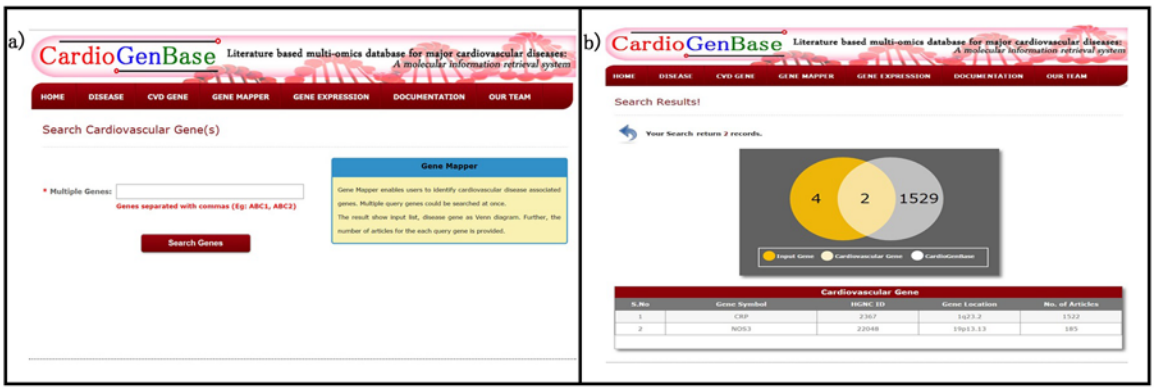

Fig 4. Gene Mapper. a) Multiple query genes can be searched at once. b) The result shows input list, disease gene as Venn diagram. Also, the number of articles for each query gene is provided.

doi:10.1371/journal.pone.0143188.g004

process. Three volunteers were assigned to collect fifty test genes associated to coronary artery disease from the articles published between the year 1988 with 2013 (Table 2). The collected genes were tested in both the databases, and their performance was validated by the volunteers. Briefly, out of fifty genes searched, most of them were present in CardioGenBase whereas only thirty six were found in CADgene database. For example, well reported coronary artery disease genes such as ALB[13], HLA[14], IL-2[15], IL-3[16], IL-27[17] and IL-33[18] were not represented with literature evidence in the CADgene database. As a result, the CardioGenBase showed better performance with respect its precision, recall, accuracy and F-measure compared to CADgene database. In addition to the performance, the volume of articles covered in CADgene is about 5000 whereas CardioGenbase contains 8319 for coronary heart disease alone. Importantly, the CardioGenbase includes literature evidence for six major cardiovascular conditions, but CADgene database is restricted only to coronary heart disease. Further, CardioGenBase provides bioavailability, gene and protein expression to aid biomarker discovery. Overall, the CardioGenBase contains more cardiovascular genes than existing databases such as CaGE[19], Phenopedia and Genopedia[20].

Conclusion and future perspectives

CardioGenBase was constructed to provide a comprehensive view of molecular information for the major cardiovascular diseases. It encompasses a broader spectrum of data by integrating

Fig 5. Gene Expression Finder. a) This tool enables users to identify gene expression in various microarray experiments associated to cardiovascular disease condition. b) the result represented as a bar diagram where the raw intensities of grouped samples are given as interactive charts.

doi:10.1371/journal.pone.0143188.g005

Table 2. List of fifty genes selected by the volunteers for validation. These fifty genes were searched in CardioGenBase and CADgene database for effective comparison. The result shows that most of the cardiac genes are found in CardioGenBase than CADgene database.

Gene Symbol	Cardiogenbase	CADgene	Volunteers
ACE	+	+	Yes
AKT1	+	-	Yes
ALB	+	-	Yes
APOC4	+	+	Yes
BCL2	+	+	Yes
BMP4	+	-	Yes
BRCA1	+	+	Yes
CASQ2	+	+	Yes
CASR	+	+	Yes
CBS	+	+	Yes
CCL11	+	+	Yes
CCL2	+	+	Yes
CMA1	+	+	Yes
CNDP1	+	+	Yes
CREG1	+	+	Yes
CRP	+	+	Yes
CSF3	+	-	Yes
CST3	+	+	Yes
EDN1	+	+	Yes
EGFR	+	+	Yes
EGR1	+	+	Yes
ENPP1	+	+	Yes
FGA	+	+	Yes
HFE	+	+	Yes
HGF	+	+	Yes
HLA-A	+	-	Yes
HLA-C	+	-	Yes
HSPB1	+	+	Yes
ICAM2	+	-	Yes
IL2	+	-	Yes
IL27	+	-	Yes
IL3	+	-	Yes
IL33	+	-	Yes
IL5	+	+	Yes
IL6	+	+	Yes
IL6R	+	+	Yes
LCN2	+	-	Yes
LDLR	+	+	Yes
LPL	+	+	Yes
MMP8	+	+	Yes
MMP9	+	+	Yes
NOS3	+	+	Yes
OCA2	-	-	No
SLC22A6	-	-	Yes
THBS4	+	+	Yes
TIMP1	+	+	Yes

(*Continued*)

Table 2. (*Continued*)

Gene Symbol	Cardiogenbase	CADgene	Volunteers
USF1	+	+	Yes
VCAM	+	+	Yes
VEGFA	+	+	Yes
XRCC3	+	+	Yes

+ and—symbol indicates presence and absence, respectively. Yes andNo indicates the cardiovascular association.

doi:10.1371/journal.pone.0143188.t002

the information from both literature and biological databases. In comparison with existing databases, CardioGenBase was created by semi-automated curation of published articles to accomplish the growing demands in the field of cardiovascular research. By providing effective search and browsing features, it operates as a flexible and user friendly platform for the molecular study of MCVDs. In the next few years, the scope of CardioGenBase will be extended to integrate new data sets with systematic updates. We hope our constant efforts would aid in understanding the molecular aspects of MCVDs that would support the global cardiovascular health.

Materials and Methods

The CardioGenBase provides extensive molecular information for the major cardiovascular diseases. The database was constructed based on (1) literature collection and curation (2) data enrichment (3) system implementation and visualization. Each of these phases is explained in the following sections.

Literature collection and curation

Gene-to-literature associations in the CardioGenBase were extracted by applying text mining approach on the records available at MEDLINE publications. In general, our approach seeks appearances of disease terms in titles, abstracts and PMC open access full text articles. Highly relevant articles were filtered and subjected to dictionary based text mining approach to extract gene/proteins. The dictionary contains both symbols as well as gene description from human gene nomenclature committee. Lucene was used to process the articles to identify gene/protein names using curated dictionary. Further, the extracted data was manually verified before data enrichment.

Data enrichment

Besides the identification of disease associated genes from the data mining, it is essential to understand their function at the molecular level. Hence, we have presented several annotations, including molecular function, biological process, cellular component, drugs, pathways, PPIs, gene and protein expression in various tissues and body fluids. Also, the bioavailability of disease-gene encoding protein is given to facilitate biomarker discovery for feasible diagnosis. All the annotation data sets were retrieved from DAVID [21],PANTHER[22], Reactome[23], HPRD[24], NCBI GEO[25], MOPED[26] and OMIM[27]. In addition, the expression profiles of these genes in various microarray datasets were provided to demonstrate their differential behavior in various cardiovascular conditions. The detail usage of the tools in database is provided in Figures A-D in S1 File.

Table 3. The parameters used validate the database. Statistics were employed to find out the precision, recall, accuracy and F-measure of CardioGenBase. Overall, the results support the viability and quality of data represented in the database.

Parameter	Cardiogenbase	CADgene
Precision	100	100
Recall	97.95	97.29
Accuracy	96.04	72.05
F-measure	98.96	98.63

doi:10.1371/journal.pone.0143188.t003

Cross validation

In order to validate the efficiency of our database, the CardioGenBase was compared with manually curated CADgene database. For reliable comparison, three volunteers were together assigned to collect fifty test genes from the research and review articles published between the years 1988 to 2013(Table 3). Further, the collected test genes were used as query to search in both the databases to determine its precision, recall, accuracy and F-measure.

$$\boldsymbol{Precision} = \frac{truepositive}{truepositive + falsepositive}$$

$$\boldsymbol{Recall} = \frac{true\ positive}{true\ positive + false\ negative}$$

$$\boldsymbol{Accuracy} = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative}$$

$$\boldsymbol{F-measure} = 2 * \frac{precision\ *\ recall}{precision + recall}$$

System implementation and visualization

A user-friendly web interface for browsing was implemented by HTML, CSS, PHP and jQuery. The data sets were stored and managed in MySQL, a popular open source database management system. All the data sets such as abstracts, ontology, gene expression, protein expression, bioavailability, pathways and drugs were maintained as separate tables. Google charts were embedded in the web page for the diagrammatic representation. In addition, jQuery, the cross-platform java script library was designed to simplify client-side scripting of HTML.

Supporting Information

S1 File. CardioGenBase tutorial for user. Describes the procedures and utility of the tools in the database.Disease Finderprovides all the genes reported for a major cardiovascular disease of interest (Figure A). CVD GENE Finderhelps the user to identify literature evidences for the gene of interest (Figure B).Gene Mapper enables users to identify cardiovascular disease associated genes. Multiple query genes could be searched at once (Figure C). Gene Expression Finder enables users to identify the gene expression in various microarray experiment associated to cardiovascular disease conditions (Figure D).
(PDF)

Acknowledgments

Authors thank Chettinad Academy of Research and Education (CARE) for computational and infrastructure facilities.

Author Contributions

Conceived and designed the experiments: AV SSSJA. Performed the experiments: AV BM DP SSSJA. Analyzed the data: AV SSSJA. Contributed reagents/materials/analysis tools: RM PGN. Wrote the paper: AV SSSJA.

References

1. Pagidipati NJ, Gaziano TA. Estimating Deaths From Cardiovascular Disease: A Review of Global Methodologies of Mortality Measurement. Circ.2013; 127: 749–756. doi: 10.1161/CIRCULATIONAHA.112.128413
2. Lee DS, Pencina MJ, Benjamin EJ, Wang TJ, Levy D, O'Donnell CJ, et al. Association of parental heart failure with risk of heart failure in offspring. N Engl J Med.2006; 355: 138–147. doi: 10.1056/NEJMoa052948 PMID: 16837677
3. Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet.2012; 44: 483–489. doi: 10.1038/ng.2232 PMID: 22446960
4. Zhang X, Johnson AD, Hendricks AE, Hwang SJ, Tanriverdi K, Ganesh SK, et al. Genetic associations with expression for genes implicated in GWAS studies for atherosclerotic cardiovascular disease and blood phenotypes. Hum Mol Genet.2014; 23: 795–782. doi: 10.1093/hmg/ddt461
5. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet.2012; 90: 7–24. doi: 10.1016/j.ajhg.2011.11.029 PMID: 22243964
6. Nanni L, Romualdi C, Maseri A, Lanfranchi G. Differential gene expression profiling in genetic and multifactorial cardiovascular diseases. J Mol Cell Cardiol.2006; 41: 934–948. doi: 10.1016/j.yjmcc.2006.08.009 PMID: 17020763
7. Sarajlić A, Janjić V, Stojković N, Radak D, Pržulj N. Network Topology Reveals Key Cardiovascular Disease Genes. PLoS One.2013; 8. doi: 10.1371/journal.pone.0071537 PMID: 23977067
8. Musunuru K, Kathiresan S. HapMap and mapping genes for cardiovascular disease. Circ Cardiovasc Genet.2008; 1: 66–71. doi: 10.1161/CIRCGENETICS.108.813675 PMID: 20031544
9. Köhler R. Single-nucleotide polymorphisms in vascular Ca2+-activated K+-channel genes and cardiovascular disease. Pflugers Arch Eur J Physiol.2010; 460: 343–351. doi: 10.1007/s00424-009-0768-6
10. Taylor IW, Wrana JL. Protein interaction networks in medicine and disease. Proteomics.2012; 12: 1706–1716. doi: 10.1002/pmic.201100594 PMID: 22593007
11. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau W-C, Ledoux P, et al. NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res.2005; 33: D562–D566. doi: 10.1093/nar/gki022 PMID: 15608262
12. Liu H, Liu W, Liao Y, Cheng L, Liu Q, Ren X,et al. CADgene: A comprehensive database for coronary artery disease genes. Nucleic Acids Res.2011; 39: D991–D996. doi: 10.1093/nar/gkq1106 PMID: 21045063
13. Sadaka M, Elhadedy A, Abdelhalim S, Elashmawy H. Albumin to creatinine ratio as a predictor to the severity of coronary artery disease. Alexandria J Med.2013; 49: 323–328. doi: 10.1016/j.ajme.2013.01.005
14. Palikhe A, Sinisalo J, Seppänen M, Valtonen V, Nieminen MS, Lokki ML. Human MHC region harbors both susceptibility and protective haplotypes for coronary artery disease. Tissue Antigens.2007; 69: 47–55. doi: 10.1111/j.1399-0039.2006.00735.x PMID: 17212707
15. Krysiak R, Okopień B. Lymphocyte-suppressing action of angiotensin-converting enzyme inhibitors in coronary artery disease patients with normal blood pressure. Pharmacol Rep.2011; 63: 1151–1161. PMID: 22180357
16. Hoffmeister A, Rothenbacher D, Bazner U, Frohlich M, Brenner H, Hombach V, et al. Role of novel markers of inflammation in patients with stable coronary heart disease. Am J Cardiol.2001; 87: 262–266. S0002-9149(00)01355-2 [pii]. PMID: 11165957
17. Jin W, Zhao Y, Yan W, Cao L, Zhang W, Wang M,et al. Elevated circulating interleukin-27 in patients with coronary artery disease is associated with dendritic cells, oxidized low-density lipoprotein, and

severity of coronary artery stenosis. Mediators Inflamm. 2012. doi: 10.1155/2012/506283 PMID: 22911112

18. Tu X, Nie S, Liao Y, Zhang H, Fan Q, Xu C,et al. The IL-33-ST2L pathway is associated with coronary artery disease in a Chinese Han population. Am J Hum Genet.2013; 93: 652–660. doi: 10.1016/j.ajhg.2013.08.009 PMID: 24075188
19. Bober M, Wiehe K, Yung C, Onal Suzek T, Lin M, Baumgartner W Jr, et al. CaGE: cardiac gene expression knowledgebase. Bioinformatics.2002; 18: 1013–1014. PMID: 12117801
20. Yu W, Clyne M, Khoury MJ, Gwinn M. Phenopedia and Genopedia: Disease-centered and Gene- centered Views of the Evolving Knowledge of Human Genetic As- sociations. Bioinformatics. 2010; 26:145–146. doi: 10.1093/bioinformatics/btp618 PMID: 19864262
21. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol.2003; 4: P3. doi: 10.1186/gb-2003-4-9-r60 PMID: 12734009
22. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. PANTHER version 7: Improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res.2009; 38: D204–D210. doi: 10.1093/nar/gkp1019 PMID: 20015972
23. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res.2014; 42: D472–D477. doi: 10.1093/nar/gkt1102 PMID: 24243840
24. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database—2009 update. Nucleic Acids Res.2009; 37: D767–D772. doi: 10.1093/nar/gkn892 PMID: 18988627
25. Acland A, Agarwala R, Barrett T, Beck J, Benson DA, Bollin C,et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res.2014; 42: D7–D17. doi: 10.1093/nar/gkt1146 PMID: 24259429
26. Kolker E, Higdon R, Haynes W, Welch D, Broomall W, Lancet D, et al. MOPED: Model Organism Protein Expression Database. Nucleic Acids Res.2012; 40: D1093–D1099. doi: 10.1093/nar/gkr1177 PMID: 22139914
27. Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). Hum Mutat.2011; 32: 564–567. doi: 10.1002/humu.21466 PMID: 21472891